1-2015

Moore's Paradox in Thought: A Critical Survey

John N. Williams
Singapore Management University, johnwilliams@smu.edu.sg

Moore's Paradox in Thought: A Critical Survey

John N. Williams, Singapore Management University

Published in Philosophy Compass, 15 January 2015, Volume10, Issue1, Pages 24-37

https://doi.org/10.1111/phc3.12188

Accepted version

Abstract

It is raining but you don't believe that it is raining. Imagine silently accepting this claim. Then you believe both that it is raining and that you don't believe that it is raining. This would be an 'absurd' thing to believe, yet what you believe might be true. It might be raining, while at the same time, you are completely ignorant of the state of the weather. But how can it be absurd of you to believe something about yourself that might be true of you? This is Moore's paradox as it occurs in thought. Solving the paradox consists in explaining why such beliefs are absurd. I give a survey of some of the main explanations. I largely deal with explanations of the absurdity of 'omissive' beliefs with contents of the form *p & I don't believe that p* and of 'commissive beliefs' with contents of the form *p & I believe that not-p* as well as beliefs with contents of the form *p & I don't know that p*.

1 The Paradox

In two different works, G.E. Moore gave the following examples of assertions:

'I went to the pictures last Tuesday but I don't believe that I did' (1942, 543)
and

'I believe that he has gone out, but he has not' (1944, 204).

Moore's first example has the 'omissive' form

p & I don't believe that *p*

so-called because it self-reports a specific lack of true belief. In contrast, his second example,

'I believe that he has gone out, but he has not'

may be formalized as

p & I believe that not-*p*.[1]

Following Sorensen (1988), this may be called the 'commissive' form, since it self-reports my specific mistake in belief. Williams (1979) was the first to draw attention to this difference, which

[1] 'I believe that he has gone out, but he has not' may be formalized as 'I believe that *p* & not-*p*'. This is equivalent to 'not-*p* & I believe that *p*', that may be represented as '*p* & I believe that not-*p*'.

stems from that between atheists and agnostics. Moore says of these utterances that '[i]t is a paradox that it should be perfectly absurd to utter assertively words of which the meaning is something which might well be true—is not a contradiction' (Baldwin 1993, 209). The paradox is the fact that these assertions are 'absurd' in some sense despite the fact that their contents might be true. Solving the paradox as it appears in speech consists in explaining why this is so. In 'Moore's Paradox in Speech: A Critical Survey', I survey some salient explanations as well as the historical emergence of the paradox.

The paradox occurs in thought as well, since if you silently believe the content of these would-be assertions then you seem no less absurd. Yet as we have just seen, the content of such an absurd belief might be true. How is that possible? What is the source of the absurdity? Must the absurdity be a form of irrationality? And why does it strike us that a contradiction is somehow at work when there is no contradiction in the content of what is believed?[2] Moore (1942, 541) also suggests that the same absurdity is found in assertions of the form

p & I don't know that *p*.

Following Williamson (2000), many agree that it would be absurd to believe such a thing (Adler 2002; Bird 2007; Huemer 2007; Sutton 2007; Unger 1975). Solving the paradox as it appears in thought consists in explaining why such beliefs are absurd.

2 Omissive and Commissive Moore-paradoxical Beliefs

Hintikka (1962) is apparently the first to deal with the paradox in thought. He claims that 'the gist of Moore's paradox may be said to lie in the fact' that the omissive proposition *p and I don't believe that p* 'is necessarily unbelievable by the speaker' (1962, 67). This is because a failure to obey *positive self-intimation*:

If one believes that *p*, then one believes that one believes that *p*[3]

'may be taken as impossible' (1962, 67).[4] So if I believe that (*p* and I don't believe that *p*), then I believe that I *don't* believe that *p*. This is because *belief distributes over conjunction*:

If one believes that (*p* & *q*), then one believes that *p* and one believes that *q*.

But by belief-distribution, I also believe that *p*, so by positive self-intimation, I believe that I *d*o believe that *p*. Thus I have contradictory second-order beliefs about whether I believe that *p*, and Hintikka seems to take this to be impossible.

[2] DeRose (1991, 59) however reports having no clear sense of inconsistency. See also Douven (2006, 475).

[3] In what follows I take the liberty of coining my own labels for principles others have used,which may or may not be their own.

[4] Negative self-intimation would be: If one doesn't believe that *p*, then one believes that one doesn't believe that *p*.

But positive self-intimation does not seem true as a universal law of psychology, given that my belief that *p* may be a prejudice that I fail to recognize within myself. Since this principle is part of an epistemic logic idealized to a perfectly rational thinker, it might be more charitable to take it as a principle of ideal rationality. But ignorance *per se* isn't irrationality, so it would need to be shown that lack of belief, and hence ignorance, of one's own mental states is an exception. Moreover, the principle involves an infinite series of beliefs about one's beliefs. It is plausible that eventually, there will occur in the series a putative belief the content of which is too complex for a human thinker to understand, or therefore believe, given *Searle's point* that one has a belief only if one has the ability to think the thought of its content (1992, 155–162). This might help explain why a failure to obey the principle seems to be a far milder fault than the irrationality of Moore-paradoxical belief.

After this, Moore-paradoxical belief seems to have received little attention, perhaps because there was little interest among epistemologists in doxastic or epistemic logic (see Hendricks and Symons 2006). But with Sorensen's *Blindspots* in 1988 came more widespread recognition that the paradox occurs in thought as well; if I silently believe

It is raining but I don't believe that it is raining

or

It is raining but I believe that it is not raining

then I seem no less absurd, yet what I believe might be true. Let us call beliefs that are possibly true yet absurd in the way Moore exemplifies, 'Moore-paradoxical' and beliefs that have the same form or syntax as these, 'Moorean'.

Sorensen made a second major contribution to the debate by observing that there are other examples of Moore-paradoxical assertions or beliefs besides Moore's own, such as

I have no beliefs now

Although you think all my opinions mistaken, you are always right

and

God knows that we are not theists.

The last two of these show that the content of Moore-paradoxical assertions or beliefs need not involve the *grammatical* singular first person, which suggests difficulties for a purely syntactic characterization of Moore-paradoxicality. Sorensen also gives the *mind-boggling case*:

The atheism of my mother's nieceless brother's only nephew angers God.

My mother's nieceless brother's only nephew can only be me. So there does seem to be a sense in which this would be an 'absurd' thing for me to believe. But since it is difficult to work through the web of relevant familial relationships, I may well be forgiven for conceiving of my mother's nieceless brother's only nephew as an existing relative *other than myself*. In that case it seems harsh to judge that my belief is irrational. This suggests that the absurdity of Moore-paradoxical assertions or beliefs might come apart from their irrationality, an important conclusion argued by Green (2007). It also suggests that the paradox requires one to conceive of oneself in first-personal terms, or in other words to have beliefs that are '*de se*'.

This suggestion appears to be corroborated by an ingenious example given by Chan (2010). Suppose that I am having a debate with John Smith on MSN Messenger, trying to convince him that the Earth is well over 5000 years old. My screen is divided into halves, labeled with my name and his. After a while, I notice that whatever I type appears on his upper half. Thinking that Smith is mimicking my words, I try to catch him out by typing

> The person actually typing these very words now here on the upper half of my screen does not believe that the Earth is well over 5,000 years old, but of course it is.

These words then appear on John Smith's screen. But unknown to me, I am the person typing the words on his screen, because the system is malfunctioning.

After Sorensen's contribution came attempts to explain the absurdity in belief (among them Baldwin (1990), de Almeida (2001), Heal (1994), Kriegel (2004) and Sorensen (2000)). Many of these again fell afoul of the difference between omissive and commissive cases (as surveyed by Green and Williams (2007), Chapter 1)

An instructive example is Heal (1994), who appeals to *positive infallibility*:

> If one believes that one believes that *p*, then one believes that *p*.

This explains the absurdity of the commissive belief: If I believe that (*p* but I believe that not-*p*), then I believe that *p*. But I also believe that I believe that not-*p*. So by positive infallibility, I believe that not-*p*. Thus I have contradictory beliefs. But this principle cannot be applied to the omissive belief. Heal could try appealing to *negative infallibility*:

> If one believes that one doesn't believe that *p* then one doesn't believe that *p*.

But applying this to omissive belief makes it impossible to have the belief, for if I believe that (*p* but I don't believe that *p*), then I do and I don't believe that *p*. Yet Heal treats Moore-paradoxical belief as an 'oddness of … thought' (1994, 6) rather than an impossible thought. Moreover it would be a strong claim indeed that it is impossible for anyone, however irrational or deranged, to hold the omissive belief.

Shoemaker (1996) approaches the paradox in terms of the *self-intimation thesis*:

> First, if a belief is available, then its subject has the belief that she has that belief, and that second-order belief is available as well. Second, if a belief is available, then if its content is presented as a candidate for assent, the subject will assent to it (1996, 218).

Shoemaker sees that the second conjunct of the first part threatens an infinite series of beliefs, but responds by restricting the thesis to cases where 'the belief that P is a first-order belief, and where … the belief that one believes that P may only be tacit' (1996, 277, note 5). The availability of a belief is, very roughly, that the belief is not repressed or stored in memory in a way that does not make it immediately apparent to one, on reflection, that one has it (1996, 217). He sums up the thesis as, 'where the subject has the concept of belief, and of herself, the first-order belief's being available *constitutes* her having the at least tacit belief that she has the first-order belief' (1996, 225).

Let us now suppose that I believe that (*p* & I don't believe that *p*) and that this belief is available to me. Assuming, as is plausible, that this means that I have an available belief in each conjunct, I have an available belief that *p*. Then by part one of the thesis, I have an available belief that I believe that *p*. So if its content is presented as a candidate for assent, I will assent to it. In this sense I am 'committed' to assenting to, 'I believe that *p*' (1996, 221). But I also have an available belief that I don't believe that *p*, so in the same sense, I am committed to assenting to 'I don't believe that *p*' as well. If I am rational, this is impossible.

Shoemaker deals with the commissive belief differently by combining the self-intimation thesis with *rational positive infallibility*:

> If one is rational and one believes that one believes that *p*, then one believes that *p*.

Suppose that I believe that (*p* & I believe that not-*p*) and that this belief is available to me. Then by part one of the thesis, I have an available belief that *p*, so by part two of the thesis, I am committed to assenting to *p*. But I also believe that I believe that not-*p*, so if I am rational, by the infallibility principle, I believe that not-*p*. Thus by part two of the thesis, I am committed to assenting to not-*p* as well. Again, if I am rational, this is impossible. Given this characterization of Shoemaker's position, objections to it by Williams (2010) appear misguided.

However, since the thesis does not apply to unavailable beliefs (as Shoemaker acknowledges, 1996, 217), this leaves unexplained the absurdity of an unavailable Moore-paradoxical belief, perhaps one that is repressed or otherwise unconscious. Moreover, one might wonder whether a *commitment* to a pair of contradictory beliefs does justice to the severe irrationality found in Moore-paradoxical belief, as opposed to *actually having them* or believing a self-contradiction, or even stronger again, believing or being conscious that one does so.

Given that a conscious belief is one of which one is aware (Lycan (2001) and Rosenthal (1986) but see Dretske (1993) for dissent), such a belief is similar to one that is available, since on reflection one immediately becomes aware that one has it. Accordingly, there is a later approach to the paradox in belief in terms of conscious belief. Its proponents include Baldwin (1990), Kriegel (2004) and Williams (2006a, 2010). Baldwin's account of the absurdity implicitly anticipates *Rosenthal's principle* (1997):

> If one consciously believes that *p*, then one believes that *p* and one believes that one oneself believes that *p*

which might be seen as a definition of conscious belief, once strengthened to a bi-conditional. The *de se* element 'one oneself' is needed in this principle. For my belief that *N* believes that *p* would not capture my awareness of my own belief even if I am *N*. I might not realize that I am *N* because I am suffering from amnesia. In that case my belief that (*p* and *N* does not believe that *p*) is not irrational. Baldwin argues that

> … a rational thinker will not consciously hold a Moorean belief. For to hold a belief consciously is both to hold the belief and be aware, and thus believe, that one holds it; and no rational thinker will believe either that he both believes and fails to believe the same thing (which is required by a conscious belief that p and that one does not believe that p) or that he both believes and disbelieves the same thing (which is required by conscious belief that p and that one believes that not-p) (1990, 230).

So if I consciously believe that (*p* & I don't believe that *p*), then I consciously believe that *p*, thus by Rosenthal's principle, I believe that I believe that *p*. Yet I also believe that I don't believe that *p*. But this result does not fit Baldwin's description of my *single* belief that I both do and don't believe that *p*. For the commissive belief, parallel reasoning delivers the result that I believe that I believe that *p* and I believe that I believe that not-*p*. This hardly counts as my single belief that I both believe and disbelieve that *p*. Deriving the single beliefs required needs the principle that *belief collects over conjunction*:

> If one believes that *p* and one believes that *q*, then one believes that (*p* & *q*).

This is dubious. I don't seem to believe the conjunction of everything I now believe, especially given Searle's point.

Kriegel (2004) deals with conscious omissive belief by appealing to the interesting *Brentano* (1874) *principle (B)*

> (B) If one consciously believes that *p*, then one believes that (*p* & one oneself believes that *p*).

This is defended by Caston (2002), Hossack (2002), Kriegel (2003a, 2003b), Smith (1986, 1989) and Thomasson (2000). Kriegel argues that a conscious omissive belief has a self-contradictory content (2004, 108–109). Since his argument is somewhat obscure, this is one occasion on which symbolism

might clarify things. Let us drop the *de se* element for ease of exposition and restrict (B) to the first-person by representing 'I believe that *p*' as 'B*p*' and 'I consciously believe that *p*' as '$B^c p$'. Then (B) is symbolized as

(B) $B^c p \rightarrow B(p \,\&\, Bp)$.[5]

Kriegel then applies (B) to the conscious omissive belief

(1) $B^c(p \,\&\, \sim Bp)$.

Substituting '$(p \,\&\, \sim Bp)$' for '*p*' in the consequent of (B) yields

(2) $B((p \,\&\, \sim Bp) \,\&\, B(p \,\&\, \sim Bp))$.

However, Kriegel represents this as

(2′) $B(p \,\&\, \sim Bp \,\&\, Bp \,\&\, \sim Bp)$.

He then invokes *Kriegel's assumption (A)* that 'believing that one believes a conjunction entails believing that one believes the conjuncts' (2004, 109), i.e.

(A) $BB(p \,\&\, q) \rightarrow B(Bp \,\&\, Bq)$.

Applying (A) to (2′), Kriegel next infers that

(3) $B(p \,\&\, B \sim p \,\&\, Bp \,\&\, B \sim Bp)$.

He observes that the content of this second-order belief is self-contradictory, since its second and third conjuncts contradict each other.

But it is unclear how (3) is supposed to follow from (A) and (2′), since (2′) contains no iterated 'BB' belief-operator. Moreover, the inference from (2) to (2′) requires not only that brackets are dropped within the scope of the second-order belief-operator but also that belief-distribution is applied within its scope. It is unclear what licenses this.

Nonetheless Kriegel can still obtain the result he wants by simply applying (B) to (1). This is because the content of the second-order belief reported by (2) is indeed self-contradictory. If $(p \,\&\, \sim Bp) \,\&\, B(p \,\&\, \sim Bp)$, then by applying &-elimination to isolate the first conjunct and then by applying &-elimination again to it, it follows that ~B*p*. But since belief-distribution applies to the second

[5] Kreigel's own symbolism is slightly different, using '$xB^C p$' to represent 'x consciously believes that p' and 'IBp' to represent 'I myself believe that p'. Nonetheless his may be easily mapped to mine in order to achieve a canonical symbolism that extends into the next note.

conjunct, it also follows that B*p*. This is a flat contradiction. So the content of an omissive belief is only a possible truth as long as one has that belief *unconsciously*.

Given this promising result, the success of Kriegel's approach will depend on how well it handles the conscious commissive belief (2004, 118, note 28). He first applies (B) to the conscious commissive belief

(4) $B^c(p \,\&\, B \sim p)$.

This yields

(5) $B((p \,\&\, B \sim p) \,\&\, B(p \,\&\, B \sim p))$.

However, Kriegel represents this as

(5′) $B(p \,\&\, B \sim p \,\&\, Bp \,\&\, B \sim p)$.

Applying (A) to (5′), Kriegel then infers that

(6) $B(p \,\&\, B \sim p \,\&\, Bp \,\&\, BB \sim p)$.

So I am guilty of believing (among other things) that I have contradictory beliefs. But it is similarly unclear how (6) is supposed to follow from (A) and (5′). We also need a plausible principle that licenses the wholesale dropping of brackets within the scope of the second-order belief-operator in (5) in order to derive to derive (5′).

This leaves us with the strategy of examining the content of the second-order belief reported by (5) in search of inconsistency. It is hard to discern. If $(p \,\&\, B \sim p) \,\&\, B(p \,\&\, B \sim p)$, then by &-elimination and belief-distribution it only follows that $Bp \,\&\, B \sim p$, and we assume that it is possible to have a pair of overtly contradictory beliefs.

Williams (2010, see also 2006a) proposes a simpler explanation by appealing to two principles. The first is that *conscious belief both distributes and collects over conjunction*:

> One consciously believes that (p & q) just in case one both consciously believes that p and one consciously believes that q

in which 'conscious' is synonymous with 'aware'.[6] The other principle is a *modification of Rosenthal's principle*:

[6] Williams (2012) supports this principle by appealing to the synchronic unity of consciousness: One is fully conscious at t of A and fully conscious at t of B just in case one is fully conscious at t of A-and-B (Bayne (2008, 2010), Tye (2003)) and the transparency of belief: If one is fully conscious at t of believing that p, then from one's point of view at t, p (Adler 2002, 196).

If one consciously has the first-order belief that *p*, then one consciously believes that one oneself believes that *p*.

The antecedent is restricted to first-order belief in order to ward off the threat of an infinite series of beliefs.

Now suppose that I become conscious of my belief that (*p* and I myself do not believe that *p*). Then since conscious belief distributes over conjunction, I have a first-order conscious belief that *p*, so by the modification of Rosenthal's principle, I have a conscious belief that I myself believe that *p*. In addition, I have a conscious belief that I myself do not believe that *p*. So since conscious belief also collects over conjunction, I have a conscious belief that I myself believe that *p* and that I myself do not believe that *p*. Thus in becoming conscious of my omissive belief, I become aware that I believe a contradiction.[7] Parallel reasoning shows that in becoming conscious of one's commissive belief, one becomes aware that one has overtly contradictory beliefs.[8]

This leaves unexplained the absurdity of an *unconscious* Moore-paradoxical belief. To deal with this, Williams appeals to *self-falsification* (2010, 242–243). Williams (1994) shows that the omissive belief is self-falsifying. All the explanations of the absurdity of Moore-paradoxical belief we have considered assume belief-distribution. A lone voice against belief-distribution is Pruss (2011, see also Williams' 2014 reply). Now suppose that I believe that (*p* and I do not believe that *p*). Then I believe that *p*. But then my second-order belief is false, since its second conjunct is false. Although my belief is not a belief in a necessary falsehood, it is self-falsifying in the sense that although what I believe might be true of me and although I might believe it, it cannot be true of me *if* I believe it. As a seeker after truth, I have shot myself in the foot. de Almeida (2001), Chan (2010) and Douven (2009) in effect observe that this does not go far enough, since analogously, there are beliefs of necessary falsehoods that are reasonable so long as one cannot be reasonably expected to recognize the necessary falsehood. But it might be said that it is indeed reasonable to expect me to recognize that

[7] For ease of exposition, we may represent this derivation symbolically. Assume that conscious belief distributes and collects over conjunction: (D&C) $(B^c p \,\&\, B^c q) \leftrightarrow B^c (p \,\&\, q)$ where '$B^c p$' represents 'I have the conscious belief that *p*', thus applying the principle to myself. Also assume the modification of Rosenthal's principle as it applies to myself, i.e. (RP′) $B_1{}^c p \rightarrow B_2{}^c {*}B_1 p$ where the numerical subscripts represent the order of belief and '*' stands for 'I myself'. The derivation now goes

1. $B^c (p \,\&\sim {*}Bp)$	Suppose
2. $B_1{}^c\, p \,\&\, B_2{}^c \sim {*}Bp$	1, D&C
3. $B_1{}^c\, p$	2, &-elimination
4. $B_2{}^c \sim {*}Bp$	2, &-elimination
5. $B_2{}^c\, {*}B_1 p$	3, RP′
6. $B_2{}^c ({*}B_1 p \,\&\sim {*}Bp)$	4,5, D&C

[8]

1. $B^c (p \,\&\sim {*}B \sim p)$	Suppose
2. $B_1{}^c\, p \,\&\, B_2{}^c\, {*}B \sim p$	1, D&C
3. $B_1{}^c\, p$	2, &-elimination
4. $B_2{}^c\, {*}B \sim p$	2, &-elimination
5. $B_2{}^c\, {*}B_1 p$	3, RP′
6. $B_2{}^c ({*}B_1 p \,\&\sim {*}B \sim p)$	4,5, D&C

my belief is self-falsifying, at least in the sense that were someone to explain to me cogently that my belief is self-falsifying, then I should accept that it is.

Turning to the commissive belief, if I believe that (*p* and I believe that not-*p*), belief-distribution again entails that I believe that *p*. But the content of the commissive belief is true only if I believe that not-*p*. Thus in having the commissive belief, its content is true only if I have overtly contradictory beliefs about whether *p*. Put another way, my commissive belief escapes self-falsification only if I have overtly contradictory beliefs, one of which must therefore be false. I have shot myself in one of two feet. Nonetheless, if I am *unaware* of having the belief, then I can hardly be expected to see that this is so, and thus can hardly be blamed for not revising my beliefs accordingly. This accounts for the intuition that any irrationality in *unconscious* Moore-paradoxical belief is far milder. Chan (2010) has objected that the self-falsification approach is circular (see Williams' 2011 reply).

Some version of the conscious belief approach might be able to explain why Moorean beliefs are not Moore-paradoxical at moments of 'dawning awareness'. Garvey (1997) gives an example of such a moment. Suppose that I am reading a self-help booklet that lists five warning signs of alcoholism. The first is that I come from a family of heavy drinkers. I think 'That's true, but I still don't believe that I am an alcoholic'. The next three signs on the list are that I occasionally drink alone, that alcohol interferes with my work and that it damages my personal life. As I read each one, I again think each time, 'That's true, but I still don't believe that I am an alcoholic'. The final sign of the list is that I do not believe that I am an alcoholic. I read this and think 'That's true' and then in dawning realization of the horrible truth, finish my thought with 'and I am an alcoholic!' (A similar example but of a prejudiced belief is given by Schwitzgebel, 2010.) This newly formed belief seems non-absurd at the *instant* it is formed although not if it is *retained*. A promising explanation of this might appeal to a theory of rational diachronic belief-revision, perhaps one consistent with theories in update semantics.

Green and Williams (2011) (following Green (2007), see also Williams (2013, 2014)) develop the self-falsification approach in terms of *norms* to explain why Moorean beliefs are normally irrational and thus absurd, while some Moorean beliefs are absurd without being irrational. Williams (2013) proposes as a norm of rational belief, the *norm of avoiding specific recognizably false beliefs*:

> Do not form—or continue to have—a specific belief that you can be reasonably expected to recognize as your very own self-falsifying belief.

Unlike in Sorensen's mind-boggling case, I normally violate this norm in having an omissive Moorean belief. So such beliefs are normally irrational, unlike Sorensen's case, which however seems absurd. Informed by Nagel (1979), Green (2007) proposes that absurdity consists in a severe violation of a system of norms such as those of belief, etiquette and conversation. One way to violate a system of norms severely is to be in a position to recognize, without further empirical investigation, that one

is doing so. However, one need not be irrational, since that violation may be very difficult to recognize. One may be in a position to recognize the violation without further empirical investigation, yet fail to actually recognize it, even if one is a genius. In contrast, one's irrationality indicates one's failure to live up to a humanly achievable standard. Building on this point, Williams (2013) holds that one norm of belief – as opposed to a norm of *rational* belief – is the *norm of avoiding false beliefs*:

Do not form—or continue to have—false beliefs.

This explains why Moorean beliefs – including that of Sorensen's mind-boggling case – are absurd, since I am in a position to see without further empirical investigation that I am violating this norm.

To explain the irrationality of the commissive belief, Williams proposes as a norm of rational belief the *norm of avoiding overtly contradictory beliefs*:

Do not form—or continue to hold—a pair of overtly contradictory beliefs.

Since the beliefs are in overt contradiction, someone who holds them can be reasonably expected to see the contradiction. As we saw above, I may only escape violating the norm of avoiding specific recognizably self-falsifying beliefs by violating this norm. This is why my commissive belief is irrational. Since I may be reasonably expected to see that my belief is self-falsifying unless I have overtly contradictory beliefs, I am *ipso facto* in a position to see this without further empirical investigation. Since one of a pair of overtly contradictory beliefs is bound to be false, I am in a position to see without further empirical investigation that I am in violation of the norm of avoiding false beliefs. So the violation is severe, and thus I am absurd.

Another interesting approach is Fernández (2005, see also 2003a and 2003b) who proposes an '*extrospective*' model of self-knowledge according to which, when one believes that *p*, then one forms the belief that one has it on the basis of the very evidence that grounds one's belief that *p*. If one has no evidence to believe that *p*, then one is justified in believing that one does not believe that *p*, as long as one's second-order belief is formed upon examining whether there is any available evidence for *p*.

Now suppose that I believe that (*p* & I don't believe that *p*). Then I believe that I don't believe that *p*. If I am justified in this, then I lack evidence for *p*, yet I also believe that *p*. If I believe that (*p* & I believe that not-*p*), then I believe that I believe that not-*p*. But if I am justified in this belief, then I have formed it on the basis of evidence for believing that not-*p*, yet I believe that *p*.

Fernández (2013, 132–138) argues that this approach avoids difficulties that afflict a related approach by Williams (2004) to which Brueckner (2006, 2009) and Vahid (2005, 2008) have also objected (see

Williams 2006b, 2009 and 2010 §3 for replies). Zimmerman (2004, 2005) and Gertler (2011) have objected to Fernández's model.

An earlier approach also in terms of evidence is given by de Almeida (2001). Let us say that a proposition *p* is *warranted* for one just in case one believes it, one has adequate evidence for it and that evidence is not 'defeated' by effective counterevidence. de Almeida suggests that we can speak of '*warrant paths* extending from the propositions that one believes to the propositions that one is entitled to believe given one's present stock of beliefs' (2001, 46). Citing Klein (1986, 266), he proposes a *Rule of Revision*:

> When a proposition *p* is added to a belief system, any belief that would block the warrant path to that belief must be removed from the belief system (2001, 47).

Now suppose that I believe that (*p* and I believe that not-*p*). Then I believe that *p* and I believe that I believe that not-*p*. This latter belief is, de Almeida claims, a reason to refrain from believing *p*. Thus I have violated the Rule of Revision. Similarly if I believe that (*p* and I do not believe *p*), then I believe that *p* and I believe that I don't believe that *p*. The fact that I believe that *p* gives me a reason for believing that I believe that *p*, which is itself reason for rejecting the belief that I don't believe that *p*. So one cannot have non-overridden evidence for a Moorean proposition. This approach may foreshadow Maitra and Weatherson's more recent discussion of 'undefeated reasons' for believing the omissive proposition (2010, 110).

Douven (2009, 365–371) appeals to principles of rational credibility. These are

Rational belief distribution (*RBD*)

> If one rationally believes that (*p* & *q*), then one rationally believes that *p* and one rationally believes that *q*

Likelihood (L)

> If one rationally believes that *p*, then one believes that it is more likely than not that *p*

Weak transparency (WT)

> If one rationally believes that *p*, then one's degree of belief that one believes that *p* is at least as great as one's degree of belief that one does not believe that *p*

and

Coherence (C)

> If one rationally believes that *p*, then it does not perspicuously follow from this, plus the principles above, that one's degrees of belief violate axioms of probability.

Now suppose for *reductio* that I rationally believe that (*p* & I don't believe that *p*) and that my degrees of belief obey the axioms of probability. By (L), my degree of belief that (*p* & I don't believe that *p*) is greater than 0.5, so by probability theory, my degree of belief that I *don't* believe that *p* is also *greater than 0.5*. But from (RBD), I rationally believe that *p*, so by (WT) my degree of belief that I *do* believe that *p* is *greater or equal to 0.5*. This violates the axioms of probability, since probability must sum to 1. So by (C) I do not rationally believe that (*p* & I don't believe that *p*) after all. It seems that this account could be extended to the commissive belief.

However it might be doubted that (L) is a principle of rational belief. Beliefs about likelihoods are relatively sophisticated. One who does not have the concept of likelihood could not obey (L) given Searle's point. But this disobedience seems to be an instance of ignorance rather than irrationality. Douven could reply that having restricted his principles to conscious *de se* belief, it is reasonable to assume this degree of conceptual sophistication. This, however, does seem to limit the scope of the explanation.

3 The Knowledge Version of the Paradox

Let us now turn to beliefs with the content *p & I don't know that p*. Suppose that I know this content. Since knowing a conjunction involves knowing each conjunct ('*knowledge-distribution*'), I know that *p*. But *knowledge is factive* (whatever is known is true). So knowing the conjunction also means that both conjuncts are true, hence I don't know that *p*. Thus I do and don't know that *p*.

Contradiction.[9] So it is impossible to know the content of the belief. Yet Williamson espouses 'the norm that one should believe that *p* only if one knows *p*' (2000, 255–256). Many defend this view (Adler 2002; Bird 2007; Huemer 2007; Sutton 2007; Unger 1975). It has been criticized by Koethe (2009) and also Littlejohn (2010) who defends the rival view that the fundamental norm of belief is truth. It might be worth noting that a weaker knowledge norm that explains the absurdity of believing *p & I don't know that p* is that one should believe only what is *possible* for one to know.

This '*knowledge account*' of assertion also accommodates the omissive belief, since its content is likewise unknowable. Suppose that I know that (*p* & I don't believe that *p*). By knowledge-distribution, I know that *p*. I also know that I don't believe that *p*, so since knowledge is factive, I

[9] Sorensen (1988) argues similarly, calling this a 'knowledge blindspot'.

don't believe that *p*. But since *knowledge entails belief* (whatever one knows, one believes), I don't know that *p*. So I do and don't know that *p*. Contradiction.

How it will handle the commissive belief is less clear. Is it impossible for me to know that (*p* & I believe that not-*p*)? An affirmative verdict might appeal to *anti-incoherence*:

If one knows that *p*, then one does not believe that not-*p*.

But could I not deceive myself into believing what I really know is false? We could restrict the principle to rational thinkers, but as a commissive variant of Garvey's example, could I not, in a fleeting instant of dawning awareness, *sensibly* come to know that I have the mistaken belief that I am not an alcoholic?[10]

To deal with the knowledge-version, Douven (2009, 371) adds to his probabilistic principles,

Weak epistemic transparency (WET)

> If one rationally believes that *p*, then one's degree of belief that one knows that *p* is at least as great as one's degree of belief that one does not know that *p*.

Now suppose that I rationally believe that (*p* & I don't know that *p*) and that my degrees of belief obey the axioms of probability. By (RBD), I rationally believe that *p*, so by (WET) my degree of belief that I *do* know that *p* is *greater or equal to 0.5*. But by (L), my degree of belief that (*p* & I don't know that *p*) is greater than 0.5, so by probability theory my degree of belief that I *don't* know that *p* is also *greater than 0.5*. This again violates axioms of probability, so by (C) I do not rationally believe that (*p* & I don't know that *p*) after all.

Given that one knows that *p* only if one has justification to believe that *p*, the absurdity should also be found in beliefs of the form *p but I have no justification to believe that p*.[11] Williams (2007, 74–75) argues that one cannot have justification to believe this. If I do, then since having justification to believe a conjunction involves having justification to believe each conjunct, I have justification to believe that *p*, so the second conjunct of my belief is false. Justification for a belief is supposed to make it a better guide to the truth. But in this case the justification guarantees a trip into falsehood and so is no justification at all.

[10] See also de Almeida's anti-incoherence: If one believes that p and also believes that not-p, then one knows neither, nor is one
justified in believing either (de Almeida 2012, 205). Suppose that I know that (p & I believe that not-p). Then since knowledge distributes over conjunction I do know that p. Thus because knowledge entails belief, I believe that p. I also know that I believe that not-p, so since knowledge is factive, I believe that not-p. Hence I believe that p and also believe that not-p, so by de Almeida's anti-incoherence, I also do not know that p. Contradiction. However the Garvey variant seems to also falsify de Almeida's anti-incoherence, since my dawning awareness seems to count as my recognition that I am an alcoholic.

[11] However, a certain kind of foundationalist might argue that there is nothing irrational in sincerely asserting 'I am now experiencing appearances of a tree but I have no justification to believe that I am now experiencing appearances of a tree (nor do I need any)'.

Another argument (Williams 2007) for the same conclusion stems from a principle that Goldman (1986, 62) seems to suggest. '*Goldman's principle*' is

If one has justification to believe that one does not have justification to believe that *p*, then one does not have justification to believe that *p*.

Now suppose that I have justification to believe that (*p* & I have no justification to believe that *p*). Then I have justification to believe that *p*. But I also have justification to believe that I do not have justification to believe that *p*, so by Goldman's principle, I do not have justification to believe that *p*. Thus I do and don't have justification to believe that *p*. Contradiction.

Goldman's principle – or principles closely related to it – has been the topic of an emerging literature on 'epistemic akrasia'; just as an akratic agent acts in a way she believes she ought not act, so an epistemically akratic subject believes something that she believes is unsupported by her evidence. Smithies (2012) argues for *Smithies principle*:

One has justification to believe that *p* just in case one has justification to believe that one has justification to believe that *p*.

He uses this to explain what is wrong with believing *p and I do not have justification to believe that p*.

Those who have argued that a rational agent cannot be epistemically akratic include Feldman (2005), Christensen (2007) and Horowitz (forthcoming). Adler (2002) argues that it is psychologically impossible to believe that *p*, be fully aware that one believes that *p*, and also believe that one's belief in *p* is unsupported by one's evidence. This idea might have been anticipated by Jones (2002), who focuses on the possibility of believing something of the form *p and the only reason I would have to believe that p is non-epistemic*. Greco (2014) argues that we should understand akratic belief states as 'fragmented', and that this justifies our taking them to be irrational.

4 Concluding Remarks

A complete solution to the paradox must accommodate all the examples we have considered, both in thought and in speech, while being mindful of the difference between omissive, commissive and knowledge cases. It must also be aware of the apparent difference between types of irrationality in assertion and belief on the one hand, and absurdity in assertion or belief on the other. It should respect the difference between conscious and unconscious belief and how this is implicated in assertion and belief-revision. This is no mean feat, but surely one worth attempting.

Biography

John N. Williams (PhD Hull) works primarily in epistemology and paradoxes, especially epistemic paradoxes. He also works in philosophy of language and applied ethics. He has published in *Acta Analytica*, *American Philosophical Quarterly*, *Analysis*, *Australasian Journal of Philosophy*, *Journal of Philosophical Research*, *Philosophy East* and *West*, *Mind, Philosophia*, *Philosophical Studies*, *Religious Studies*, *Social Epistemology Review and Reply Collective*, *Synthese* and *Theoria.* He is co-editor of *Moore's Paradox: New Essays on Belief, Rationality and the First Person,* Oxford University Press together with Mitchell Green. He researches and teaches in the School of Social Sciences, Singapore Management University.

Works Cited

Adler, J. Belief's Own Ethics. Cambridge, MA: MIT University Press, 2002.

Baldwin, T. G.E. Moore. London: Routledge, 1990.

Baldwin, T. G.E. Moore: Selected Writings. London: Routledge, 1993.

Bayne, T. ' The Unity of Consciousness and the Split Brain Syndrome.' Journal of Philosophy 105 (2008): 277– 300.

Bayne, T. The Unity of Consciousness. Oxford: Oxford University Press, 2010.

Bird, A. ' Justified Judging.' Philosophy and Phenomenological Research 74 (2007): 81– 110.

Brentano, F. Psychology from an Empirical Standpoint. 1874. Trans. A. C. Rancurello, D. B. Terrell, and L. L. McAlister. London: Routledge and Kegan Paul, 1973.

Brueckner, A. ' Justification and Moore's Paradox.' Analysis 66 (2006): 264– 6.

Brueckner, A. ' More on Justification and Moore's Paradox.' Analysis 69 (2009): 497– 9.

Caston, V. ' Aristotle on Consciousness.' Mind 111 (2002): 751– 815.

Chan, T. ' Moore's Paradox is not Just Another Pragmatic Paradox.' Synthese 173 (2010): 211– 29.

Christensen, D. ' Does Murphy's Law Apply in Epistemology? Self-Doubt and Rational Ideals.' Oxford Studies in Epistemology 2 (2007): 3– 31.

de Almeida, C. ' What Moore's Paradox is About.' Philosophy and Phenomenological Research 62 (2001): 33– 58.

de Almeida, C. ' Epistemic Closure, Skepticism and Defeasibility.' Synthese 188 (2012): 197– 215.

DeRose, K. ' Epistemic Possibilities.' Philosophical Review 100 (1991): 581– 605.

Douven, I. ' Assertion, Knowledge, and Rational Credibility.' The Philosophical Review 115 (2006): 449– 85.

Douven, I. ' Assertion, Moore, and Bayes.' Philosophical Studies 144 (2009): 361– 75.

Dretske, F. I. ' Conscious Experience.' Mind 102 (1993): 263– 83.

Feldman, R. ' Respecting the Evidence.' Philosophical Perspectives 19 (2005): 95– 119.

Fernández, J. ' Privileged Access Naturalized.' Philosophical Quarterly 53 (2003a): 352– 72.

Fernández, J. ' Privileged Access Revisited.' Philosophical Quarterly 55 (2003b): 102– 5.

Fernández, J. ‘ Self-Knowledge, Rationality and Moore’s Paradox.’ Philosophy and Phenomenological Research 71 (2005): 533– 56.

Fernández, J. Transparent Minds: A Study of Self-Knowledge. Oxford: Oxford University Press, 2013.

Garvey, J. ‘ Believing P but Not P.’ Cogito 11 (1997): 14– 6.

Gertler, B. ‘ Self-Knowledge and the Transparency of Belief.’ Self-Knowledge. Ed. A. Hatzimoysis. Oxford: Oxford University Press, 2011. 125– 45.

Goldman, A. Epistemology and Cognition. Cambridge, Mass.: Harvard University Press, 1986.

Greco, D. ‘ A Puzzle About Epistemic Akrasia.’ Philosophical Studies 167 (2014): 201– 19.

Green, M. S. ‘ Moorean Absurdity and Showing What’s Within.’ Moore’s Paradox: New Essays on Belief, Rationality and the First Person. Eds. M. S. Green and J. N. Williams. Oxford: Oxford University Press, 2007. 189– 214.

M. S. Green and J. N. Williams, eds. Moore’s Paradox: New Essays on Belief, Rationality and the First Person. Oxford: Oxford University Press, 2007.

Green, M. S. and J. N. Williams. ‘ Moore’s Paradox, Truth and Accuracy.’ Acta Analytica 26 (2011): 243– 55.

Heal, J. ‘ Moore’s Paradox: A Wittgensteinian Approach.’ Mind 103 (1994): 5– 24.

Hendricks, V. and J. Symons. ‘ Where’s the Bridge? Epistemology and Epistemic Logic.’ Philosophical Studies 128 (2006): 137– 67.

Hintikka, J. Knowledge and Belief. Ithaca, NY: Cornell University Press, 1962.

Horowitz, S. ‘ Epistemic Akrasia.’ Noûs (forthcoming).

Hossack, K. ‘ Self-Knowledge and Consciousness.’ Proceedings of the Aristotelian Society 102 (2002): 163– 81.

Huemer, M. ‘ Moore’s Paradox and the Norm of Belief.’ Themes from G. E. Moore: New Essays in Epistemology and Ethics. Eds. S. Nuccetelli and G. Seay. Oxford: Oxford University Press, 2007. 142– 57.

Jones, W. ‘ Explaining Our Own Beliefs: Non-Epistemic Believing and Doxastic Instability.’ Philosophical Studies 111 (2002): 217– 49.

Klein, P. ‘ Immune Belief Systems.’ Philosophical Topics 14 (1986): 259– 80.

Koethe, J. ‘ Knowledge and the Norms of Assertion.’ Australasian Journal of Philosophy 87 (2009): 625– 38.

Kriegel, U. ‘ Consciousness as Intransitive Self-Consciousness: Two Views and an Argument.’ Canadian Journal of Philosophy 33 (2003a): 103– 32.

Kriegel, U. ‘ Consciousness, Higher-Order Content, and the Individuation of Vehicles.’ Synthese 134 (2003b): 477– 504.

Kriegel, U. ‘ Moore’s Paradox and the Structure of Conscious Belief.’ Erkenntnis 61 (2004): 99– 121.

Littlejohn, C. ‘ Moore’s Paradox and Epistemic Norms.’ Australasian Journal of Philosophy 88 (2010): 79–100.

Lycan, W. G. ‘ A Simple Argument for a Higher-Order Representation Theory of Consciousness.’ Analysis 61 (2001): 3– 4.

Maitra, I. and B. Weatherson. ‘ Assertion, Knowledge and Action.’ Philosophical Studies 149 (2010): 99– 118.

Moore, G. E. ‘ A Reply to My Critics.’ The Philosophy of G. E. Moore. Ed. P. Schilpp. La Salle, Ill: Open Court, 1942. 535– 677.

Moore, G. E. ‘ Russell’s Theory of Descriptions.’ The Philosophy of Bertrand Russell. Ed. P. Schilpp. La Salle, Ill: Open Court, 1944. 175– 225.

Nagel, T. ‘ The Absurd.’ Mortal Questions. Ed. T. Nagel. Cambridge: Cambridge University Press, 1979. 11– 24.

Pruss, A. Does belief distribute over conjunction? 2011. 19 July 2012 <http://alexanderpruss.blogspot.sg/2011/10/does-belief-distribute-over-conjunction.html>.

Rosenthal, D. M. ‘ Two Concepts of Consciousness.’ Philosophical Studies 94 (1986): 329– 59.

Rosenthal, D. M. ‘ A Theory of Consciousness.’ The Nature of Consciousness: Philosophical Debates. Eds. N. J. Block, O. Flanagan and G. Güzeldere. Cambridge, MA: MIT Press and Bradford Books, 1997. 729– 53.

Schwitzgebel, E. ‘ Acting Contrary to Our Professed Beliefs.’ Pacific Philosophical Quarterly 91 (2010): 531– 53.

Searle, J. The Rediscovery of the Mind. Cambridge, MA: MIT Press, 1992.

Shoemaker, S. ‘ Moore’s Paradox and Self-knowledge.’ Philosophical Studies 77 (1996): 211– 28.

Smith, D. W. ‘ The Structure of (Self-) Consciousness.’ Topoi 5 (1986): 149– 56.

Smith, D. W. The Circle of Acquaintance. Dordrecht: Kluwer, 1989.

Smithies, D. ‘ Moore’s Paradox and the Accessibility of Justification.’ Philosophy and Phenomenological Research 85 (2012): 273– 300.

Sorensen, R. A. Blindspots. Oxford: Clarendon Press, 1988.

Sorensen, R. A. ‘ Moore’s Problem with Iterated Belief.’ Philosophical Quarterly 50 (2000): 28– 43.

Sutton, J. Without Justification. Cambridge, MA: MIT University Press, 2007.

Thomasson, A. L. ‘ After Brentano: A One-Level Theory of Consciousness.’ European Journal of Philosophy 8 (2000): 190– 209.

Tye, M. Consciousness and Persons: Unity and Identity. Cambridge, MA: MIT University Press, 2003.

Unger, P. Ignorance: The Case for Skepticism. Oxford: Clarendon Press, 1975.

Vahid, H. ‘ Moore’s Paradox and Evans’s Principle: A Reply to Williams.’ Analysis 65 (2005): 337– 41.

Vahid, H. ‘ Radical Interpretation and Moore’s Paradox.’ Theoria 74 (2008): 146– 63.

Williams, J. N. ‘ Moore’s Paradox – One or Two?’ Analysis 39 (1979): 141– 2.

Williams, J. N. ‘ Moorean Absurdity and the Intentional “Structure” of Assertion.’ Analysis 54 (1994): 160– 6.

Williams, J. N. ‘ Moore’s Paradoxes, Evans’s Principle and Self-Knowledge.’ Analysis 64 (2004): 348– 53.

Williams, J. N. ‘ Moore’s Paradoxes and Conscious Belief.’ Philosophical Studies 127 (2006a): 383–414.

Williams, J. N. ‘ In Defence of an Argument for Evans’s Principle: A Rejoinder to Vahid.’ Analysis 66 (2006b): 167– 70.

Williams, J. N. ‘ The Surprise Exam Paradox: Disentangling Two Reductios.’ Journal of Philosophical Research 32 (2007): 67– 95.

Williams, J. N. ‘ Justifying Circumstances and Moore-paradoxical Beliefs: A Response to Brueckner.’ Analysis 69 (2009): 490– 6.

Williams, J. N. ‘ Moore’s Paradox, Defective Interpretation, Justified Belief and Conscious Belief.’ Theoria 76 (2010): 211– 48.

Williams, J. N. ‘ The Completeness of the Pragmatic Solution to Moore’s Paradox in Belief: A Reply to Chan.’ Synthese 190 (2011): 2457– 2476.

Williams, J. N. ‘ Moore-paradoxical Assertion, Fully Conscious Belief and the Transparency of Belief.’ Acta Analytica 27 (2012): 9– 12.

Williams, J. N. ‘ Moore’s Paradox and the Priority of Belief Thesis.’ Philosophical Studies 165 (2013): 1117– 38.

Williams, J. N. ‘ Moore’s Paradox in Belief and Desire.’ Acta Analytica 29 (2014): 1– 23.

Williamson, T. Knowledge and Its Limits. Oxford: Oxford University Press, 2000.

Zimmerman, A. ‘ Unnatural Access.’ Philosophical Quarterly 54 (2004): 435– 8.

Zimmerman, A. ‘ Putting Extrospection to Rest.’ Philosophical Quarterly 55 (2005): 658– 61.